CSE

Code: **20ACS29**

R20

III B.Tech II Semester Regular Examinations, July/August 2023

Data Warehousing and Data Mining

(Common to CSE, CSE(DS) & IT)

(For 2020 Admitted Batch Only)

Time: 3 Hours Max.Marks:60

*****

PART-A

Answer all questions and all questions carry equal marks 5 x 2 = 10M

1. List out any Two (2) differences between discrete and continuous attributes with examples. (CO1 - Understand)
2. Give the Formulae for Pearson's product moment coefficient and explain the terms in it. (CO2 - Remember)
3. There are Three (3) measures which return good results while comparing attribute selection members. Name them. (CO3 - Understand)
4. Cluster Analysis is having applications in "Biology" & "Climate". Give your views on it. (CO4 - Remember)
5. Deduct the challenges of web in knowledge discovery. (CO5 - Analyze)

PART-B

Answer All questions and all questions carry equal marks 5 x 10 = 50M

6 a) Categorize the steps involved in knowledge discovery in databases. (CO1 - Remember)

(OR)

b) Classify OLAP Operations. (CO1 - Understand)

7 a) Summarize in detail about various kinds of association rules. (CO2 - Understand)

(OR)

b) Analyze the various Frequent Itemset mining method with examples. (CO2 - Analyze)

8 a) Generalize the Bayes theorem of posterior probability and explain the working of a Bayesian classifier with an example. (CO3 - Understand)

(OR)

b) Formulate Rule Based Classification Techniques. (CO3 - Analyze)

9 a) Organize your views on Cluster Analysis and intercept the dissimilarity measures for interval scaled variables and binary variables. (CO4 - Understand)

(OR)

b) Summarize basic Clustering Methods. (CO4 - Understand)

10 a) *"Suppose that a text retrieval system has just retrieved a number of documents for me based on my input in the form of a query"*. How can we access how accurate or correct the system was? Sketch the relationship of the above statement. (CO5 - Remember)

(OR)

b) Summarize Text Indexing techniques. (CO5 - Understand)

R20

III B. Tech II Semester Regular Examinations July / August - 2023

Data Warehousing and Data Mining

Subject Code: 20ACS29

Branch: Common to CSE, CSE (DS) & IT **Max Marks: 60**

(For 2020 Admitted Batch Only)

PART - A
Answer All Questions and all questions carry equal marks (5*2=10)

Q.NO	Key Points for valuation	Total marks
1.	List out any Two (2) differences between discrete and continuous attributes with examples. Discrete Data - Bar graphs are a visual representation of discrete data. Discrete data can be counted. Continuous Data - Continuous data is quantifiable. Continuous data are graphically represented using a histogram.	2
2.	Give the Formulae for Pearson's product moment coefficient and explain the terms in it. The Pearson product-moment correlation coefficient formula is: $r = \frac{N\Sigma xy-(\Sigma x)(\Sigma y)}{\sqrt{[N\Sigma x^2-(\Sigma x)^2][N\Sigma y^2-(\Sigma y)^2]}}$ The terms in that formula are: N = the number of data points, i.e., (x, y) pairs, in the data set.	2
3.	There are Three (3) measures which return good results while comparing attribute selection members. Name them. There are three popular attribute selection measures: Information Gain, Gain ratio, and, Gini index.	2
4.	Cluster Analysis is having applications in "Biology" & "Climate". Give your views on it. Cluster analysis is the name given to a set of techniques which ask whether data can be grouped into categories on the basis of their similarities or differences. It began when biologists started to classify plants on the basis of their various phyla and species and wanted to derive a less subjective technique. Clustering techniques are used in the analysis of weather and climate to identify distinct, discrete groups of atmospheric and oceanic structures and evolutions from observations, reanalyzes, and numerical model simulations and predictions.	2
5.	Deduct the challenges of web in knowledge discovery. These challenges include: Missing values; Data scarcity, Data dimensionality reduction, Black box; and Mathematical model.	2

	PART - B **Answer All Questions and all questions carry equal marks (5*10=50)**	
6.a)	Categorize the steps involved in knowledge discovery in databases. The KDD Process Steps: 1. Building up an understanding of the application domain. ADVERTISEMENT. 2. Choosing and creating a data set on which discovery will be performed. 3. Preprocessing and cleansing. 4. Data Transformation. 5. Prediction and description. 6. Selecting the Data Mining algorithm. 7. Utilizing the Data Mining algorithm. 8. Evaluation. 9. Using the discovered knowledge.	**10**
6.b)	Classify OLAP operations. OLAP is considered (Online Analytical Processing) which is a type of software that helps in analyzing information from multiple databases at a particular time. OLAP is simply a multidimensional data model and also applies querying to it. OLAP operations: There are five basic analytical operations that can be performed on an OLAP cube: 1. Drill down 2. Roll up 3. Dice (Topic with explanation) 4. Slice 5. Pivot	**10**
7.a)	Summarize in detail about various kinds of association rules. Association rule learning is a machine learning technique used for discovering interesting relationships between variables in large databases. It is designed to detect strong rules in the database based on some interesting metrics. Types of Association Rules: There are various types of association rules in data mining: - Multi-relational association rules Generalized association rules (Topic with explanation) Quantitative association rules Interval information association rules	**10**
7.b)	Analyze the various Frequent Itemset mining method with examples. Frequent item sets, also known as association rules, are a fundamental concept in association rule mining, which is a technique used in data mining to discover relationships between items in a dataset. Frequent Itemsets are determined by Apriori, Eclat, and FP-growth algorithms. Apriori algorithm is the commonly used frequent itemset mining algorithm. It works well for association rule learning over transactional and relational databases	

	Example On finding Frequent Itemsets – Consider the given dataset with given transactions. | TransactionId	Items
---	---
1	{A,C,D}
2	{B,C,D}
3	{A,B,C,D}
4	{B,D}

5	{A,B,C,D} |
	
	Lets say minimum support count is 3
	Relation hold is maximal frequent => closed => frequent
	1-frequent: {A} = 3; // not closed due to {A, C} and not maximal {B} = 4; // not closed due to {B, D} and no maximal {C} = 4; // not closed due to {C, D} not maximal {D} = 5; // closed item-set since not immediate super-set has same count. Not maximal
	2-frequent: {A, B} = 2 // not frequent because support count < minimum support count so ignore {A, C} = 3 // not closed due to {A, C, D} {A, D} = 3 // not closed due to {A, C, D} {B, C} = 3 // not closed due to {B, C, D} {B, D} = 4 // closed but not maximal due to {B, C, D} {C, D} = 4 // closed but not maximal due to {B, C, D}
	3-frequent: {A, B, C} = 2 // ignore not frequent because support count < minimum support count {A, B, D} = 2 // ignore not frequent because support count < minimum support count {A, C, D} = 3 // maximal frequent {B, C, D} = 3 // maximal frequent
	4-frequent: {A, B, C, D} = 2 //ignore not frequent </
		10

8.a)	Generalize the Bayes theorem of posterior probability and explain the working of Bayesian classifier with an example. Bayes' Theorem describes the probability of an event, based on precedent knowledge of conditions which might be related to the event. In other words, Bayes' Theorem is the add-on of Conditional Probability. With the help of Conditional Probability, one can find out the probability of X given H, and it is denoted by P(X | H). Now Bayes' Theorem states that if we know Conditional Probability (P(X | H)) then we can find out P(H | X), given the condition that P(X) and P(H) are already known to us. Bayes' Theorem is named after Thomas Bayes. He first makes use of conditional probability to provide an algorithm which uses evidence to calculate limits on an unknown parameter. Bayes' Theorem has two types of probabilities : Prior Probability [P(H)] Posterior Probability [P(H/X)] Working of Naïve Bayes' Classifier can be understood with the help of the below example: Suppose we have a dataset of weather conditions and corresponding target variable "Play". So using this dataset we need to decide that whether we should play or not on a particular day according to the weather conditions.	**10**
8.b)	Formulate Rule Based Classification Techniques. IF-THEN Rules	

	Rule-based classifier makes use of a set of IF-THEN rules for classification. We can express a rule in the following from – Let us consider a rule R1, R1: IF age = youth AND student = yes THEN buy_computer = yes Points to remember – The IF part of the rule is called rule antecedent or precondition. The THEN part of the rule is called rule consequent. The antecedent part the condition consist of one or more attribute tests and these tests are logically ANDed. The consequent part consists of class prediction. Note – We can also write rule R1 as follows – R1: (age = youth) ^ (student = yes))(buys computer = yes) If the condition holds true for a given tuple, then the antecedent is satisfied.	**10**
9.a)	Organize your views on Cluster Analysis and intercept the dissimilarity measures for interval scaled variables and binary variables. • Cluster analysis is the grouping of objects such that objects in the same cluster are more similar to each other than they are to objects in another cluster. • The classification into clusters is done using criteria such as smallest distances, density of data points, graphs, or various statistical distributions. • Interval-scaled variables are continuous data of an approximately linear scale. • An examples such as weight and height, latitude and longitude coordinates (e.g., when clustering homes), and weather temperature. • The measurement unit used can influence the clustering analysis. • Similarity or distance measures are core components used by distance-based clustering algorithms to cluster similar data points into the same clusters, while dissimilar or distant data points are placed into different clusters.	**10**
9.b)	Summarize basic Clustering methods. Clustering methods can be classified into the following categories – Partitioning Method Hierarchical Method Density-based Method (Topics with explanation) Grid-Based Method Model-Based Method Constraint-based Method	**10**
10.a)	"*Suppose that a text retrieval system has just retrieved a number of documents for me based on my input in the form of a query*". How can we access how accurate or correct the system was? Sketch the relationship of the above statement. Let the set of documents relevant to a query be denoted as [Relevant], and the set of documents retrieved be denoted as [Retrieved].	

	There are two basic measures for assessing the quality of text retrieval: **Precision:** This is the percentage of retrieved documents that are in fact relevant to the query (i.e., "correct" responses). It is formally defined as Precision = |{ Retrieved} intersection { Relevant }|/|{Retrieved}| **Recall:** This is the percentage of documents that are relevant to the query and were in fact, retrieved. It is formally defined as Recall = {Relevant} intersection {Retrieved}/{Retrieved} Most information retrieval systems support keyword-based and/or similarity-based retrieval. In keyword- based information retrieval, a document is represented by a string, which can be identified by a set of keywords. A user provides a keyword or an expression formed out of a set of keywords, such as "car and repair shops", "tea or coffee", or "database systems but not Oracle". A good information retrieval system should consider synonyms when answering such queries.	**10**
10.b)	Summarize Text Indexing techniques. There are several popular text retrievals indexing techniques such as inverted indices and signature files. ✓ Inverted Index - An inverted index is an index structure that maintains two hash indexed or B+-tree indexed tables: document_table and term_table, where document_table consists of a set of document records, each including two fields: doc_id and posting_list, where posting_list is a list of methods (or pointers to methods) that appears in the document, arranged according to some relevance measure. ✓ The retrieval system is often required to return relevant document/text for a user query in a few seconds. ✓ Without an index, it is impossible for a retrieval system to achieve this task. ✓ There are many indexing techniques. Among them, • inverted index, • suffix array, and • signature are three typical examples. Topics with explanation	**10**

Prepared By,

Mr. P. Nandakumar
Assistant Professor,
Department of IT,
SVCET, Chittoor.
Mobile: 9894388230/
8667226163